Ph. D. Final Synopsis

On

Role of AI & ML in Prognosis and Profiling of Breast Cancer by Single-Cell RNA Sequencing (sc-RNA-seq)

Submitted to the

MGM University, Aurangabad (MS)- India.

For the Completion of Ph.D. Degree
In
Computer Science and Information Technology

Submitted by

Ms. Ruheena Hashmi Syed Mubeen Hashmi
Research Scholar
MGM University, Chhatrapati Sambhaji Nagar

Under the Guidance

Dr. Dnyaneshwari D. Patil
Assistant Professor/Guide
Dr. G. Y. Pathrikar College of CS and IT.
MGM University, Chhatrapati Sambhaji Nagar

Index

Sr. No.	Table of Contents	Pages
	Abstract	3
1.	Introduction	4
2.	Literature Review	5
3.	Definition of the Problem	7
4.	Objectives	8
5.	Significance of the study	8
6.	Major Contribution	8
7.	Research Processing Steps	8
8.	Research Methods and Material	12
9.	Results and Discussion	16
10.	Conclusion	29
11.	Chapter Scheme	30
12.	Research papers Presented and Published.	31
13.	References.	33

ABSTRACT

Breast cancer is one of the most prevalent cancers and a leading cause of cancer-related deaths among women globally. It represents a major public health challenge due to its complex heterogeneity and the associated difficulties in early detection and accurate subtype classification. Despite advancements in imaging-based diagnostics such as mammography and MRI, these techniques often suffer from significant limitations, including high rates of false positives and false negatives. Additionally, traditional approaches fail to capture the molecular diversity of tumors, limiting their utility for precision medicine and targeted therapy.

This study aims to overcome these challenges by integrating cutting-edge technologies: **Single-Cell RNA Sequencing (scRNA-seq)** and **Artificial Intelligence (AI)** with **Machine Learning (ML)**. scRNA-seq enables the examination of gene expression at the single-cell level, providing a high-resolution view of cellular heterogeneity within tumors. This study utilizes scRNA-seq data to identify molecular biomarkers critical to breast cancer progression, such as **BRCA1**, **PTEN**, and **ESR1**. These genes are not only pivotal in tumor biology but also hold potential for improving diagnosis, prognosis, and personalized treatment strategies.

To analyze this high-dimensional scRNA-seq data, advanced AI/ML models were developed and validated. Algorithms such as **Random Forest (RF)**, **Support Vector Machine (SVM)**, **Decision Tree (DT)**, and **K-Nearest Neighbors (KNN)** were applied to datasets sourced from publicly available genomic repositories, such as GEO (e.g., GSE118389, GSE161529). The data underwent rigorous preprocessing, including normalization, feature selection (e.g., identification of highly variable genes), and dimensionality reduction techniques (e.g., PCA, t-SNE). These methods ensured the robustness and accuracy of downstream analysis.

Key findings of the study include:

1. **Subtype Classification:** Random Forest achieved the highest classification accuracy of **98.16%**, demonstrating its superiority in handling complex and imbalanced datasets.
2. **Malignant vs. Benign Differentiation:** Significant differences in the expression levels of genes such as BRCA1 between malignant and benign cells were identified, highlighting their potential as diagnostic biomarkers.
3. **Treatment Personalization:** By identifying subtype-specific biomarkers, the study lays the foundation for precision medicine approaches, enabling the tailoring of therapies to individual tumor profiles.

This research underscores the transformative potential of integrating scRNA-seq with AI/ML in oncology. It not only improves diagnostic accuracy and subtype classification but also offers scalable solutions for future research and clinical applications. The framework developed in this study paves the way for further innovations

in precision oncology, particularly in developing targeted treatments and predictive models for breast cancer outcomes.

The findings represent a significant step forward in leveraging computational biology for breast cancer research, offering a pathway toward improving patient outcomes and advancing the field of precision medicine.

1. INTRODUCTION

Breast cancer is one of the most common cancers among women and a leading cause of cancer-related deaths worldwide. Accurate diagnosis and effective treatment planning are critical for improving patient outcomes. Traditionally, breast cancer diagnostics have relied on imaging techniques such as **mammography** and **magnetic resonance imaging (MRI)**. These methods have proven effective for detecting tumors and reducing mortality rates; however, they face several limitations, particularly in cases involving dense breast tissue. Dense breast tissue can obscure tumors on mammograms, leading to false negatives, while benign abnormalities may appear suspicious, resulting in false positives. These inaccuracies increase patient anxiety, delay proper treatment, and add to healthcare costs.

Recent advancements in **molecular profiling techniques**, particularly **Single-Cell RNA Sequencing (scRNA-seq)**, have revolutionized cancer research by providing a deeper understanding of tumor biology. Unlike traditional bulk sequencing, which averages gene expression across a large population of cells, scRNA-seq enables the analysis of gene expression at the single-cell level. This high-resolution technology allows researchers to uncover tumor heterogeneity, identify rare cell populations, and investigate the tumor microenvironment in detail. For instance, scRNA-seq can reveal how specific genes, such as **BRCA1, PTEN, and ESR1**, contribute to cancer progression and subtype differentiation. Such insights are essential for the development of personalized treatment strategies.

However, the potential of scRNA-seq data is often underutilized due to its **high-dimensional nature**, which poses computational challenges. Each dataset may contain thousands of cells and tens of thousands of genes, leading to complexities in analysis and interpretation. Sophisticated computational approaches are required to extract meaningful patterns from this data. This is where **Artificial Intelligence (AI)** and **Machine Learning (ML)** play a transformative role.

AI/ML techniques are well-suited for handling large-scale, high-dimensional datasets. They can analyze complex patterns, predict outcomes, and identify biomarkers with unparalleled precision. For example:

- **Supervised learning models** like Random Forest (RF) and Support Vector Machine (SVM) can classify breast cancer subtypes based on gene expression profiles.

- **Unsupervised learning methods**, such as clustering algorithms, can identify novel cell populations and tumor characteristics.
- Predictive models can assess the likelihood of disease progression or response to specific therapies.

This study combines the strengths of **scRNA-seq** and **AI/ML** to address the limitations of traditional diagnostic methods. By analyzing scRNA-seq data, the research aims to:

1. Improve the accuracy of breast cancer diagnostics.
2. Enhance the classification of subtypes such as Basal, Luminal A, Luminal B, and Her2.
3. Provide a framework for personalized treatment strategies by identifying molecular biomarkers.

The integration of AI/ML with scRNA-seq not only improves diagnostic precision but also supports the development of innovative tools for precision oncology. By leveraging cutting-edge technology, this research seeks to bridge the gap between molecular biology and clinical applications, ultimately improving outcomes for breast cancer patients.

2. LITERATURE REVIEW

2.1. Advancements in Breast Cancer Diagnostics

Imaging-based diagnostics such as mammography and MRI have significantly reduced breast cancer mortality by up to 30%, particularly in women aged 50-69 years (Duffy et al., 2020). While effective, these methods face limitations:

- **False Positives/Negatives:** Mammograms yield false positives (~10%), leading to unnecessary interventions, while dense breast tissue often results in false negatives (American Cancer Society, 2023).
- **Invasiveness:** Biopsies, required for confirming diagnoses, are associated with discomfort and potential complications.

These limitations underscore the need for more precise, non-invasive methods that can address the heterogeneity of breast tumors.

2.2. Single-Cell RNA Sequencing (scRNA-seq)

The advent of scRNA-seq has revolutionized breast cancer research by providing cellular-level resolution of gene expression, surpassing the limitations of bulk RNA sequencing (La Manno et al., 2018).

Its applications include:

- **High-Resolution Profiling:**scRNA-seq captures tumor heterogeneity and reveals distinct subpopulations such as cancer stem cells and immune cells.

- **Biomarker Discovery:** Genes like BRCA1, PTEN, and ESR1 have been identified as critical biomarkers for diagnosis and prognosis.
- **Tumor Microenvironment Insights:**scRNA-seq elucidates cancer-cell interactions with their microenvironment, aiding in understanding metastasis and treatment resistance (Patel et al., 2014).

Despite its advantages, the high dimensionality of scRNA-seq data requires advanced computational tools for analysis.

2.3. Applications of AI/ML in Breast Cancer Research

AI/ML techniques have shown promise in addressing the complexities of high-dimensional genomic data:

- **Random Forest (RF):** Achieves high accuracy in classifying breast cancer subtypes by aggregating predictions from decision trees (Goodfellow et al., 2016).
- **Support Vector Machine (SVM):** Effectively classifies malignant vs. benign cells, even in small, high-dimensional datasets (McKinney et al., 2020).
- Unsupervised Learning: Techniques like clustering uncover novel tumor subpopulations, enhancing subtype classification (Kiselev et al., 2019).

These methods enable biomarker discovery, subtype classification, and prognosis prediction, paving the way for precision oncology.

2.4 Existing Gaps in Research

Despite the potential of scRNA-seq and AI/ML, their integration remains underexplored:

Few studies have applied AI/ML to analyze scRNA-seq data for breast cancer diagnostics and prognosis (La Manno et al., 2018).

Challenges such as data dimensionality, integration into clinical workflows, and lack of generalizable models hinder broader applications (Stegle et al., 2015).

While advancements in imaging and molecular profiling have improved breast cancer diagnostics, limitations persist in addressing tumor heterogeneity and precision medicine. The integration of scRNA-seq with AI/ML offers a promising solution to these gaps. This study bridges this critical research void by leveraging advanced computational methods for breast cancer profiling, enabling accurate diagnostics, subtype classification, and personalized treatments.

2.5 Proposed Solution

This research aims to bridge the gap by developing an AI/ML framework tailored for sc-RNA-seq data analysis. Key objectives include:

- **Early Detection**: Leveraging AI/ML to identify molecular signatures of cancer at the earliest stages.

- **Subtype Classification:** Building robust ML models for accurately distinguishing breast cancer subtypes.
- **Prognosis and Personalized Treatment:** Using AI/ML to predict treatment responses and inform personalized therapy strategies.
- **Clinical Integration:** Creating a scalable, user-friendly diagnostic tool for use in real-world clinical settings.

The proposed framework addresses the computational challenges of scRNA-seq while unlocking its potential to transform breast cancer diagnostics and treatment personalization.

3. DEFINITION OF THE PROBLEM

Breast cancer remains a major global health challenge despite advances in diagnostic technologies and treatment. While imaging techniques like mammography, ultrasound, and MRI have improved survival rates, they still face limitations in early detection, accurate subtype classification, and personalized treatment, affecting patient outcomes.

Early Detection Gaps

Early detection is critical for successful treatment, but current imaging methods have limitations. Mammography, for instance, misses early-stage tumors in women with dense breast tissue, delaying diagnosis and limiting treatment options. False positives also lead to unnecessary biopsies and increased healthcare costs. Imaging is less effective at detecting small, early-stage tumors, underscoring the need for more sensitive diagnostic tools.

Molecular Profiling Limitations

Traditional diagnostic methods fail to fully capture the molecular and genetic diversity of breast cancer. Bulk RNA sequencing, while useful, provides an averaged view of tumor samples, missing critical information about rare cell populations and tumor heterogeneity. This limits the ability to develop personalized treatments based on specific tumor subtypes, resulting in generalized treatment plans.

Computational Bottlenecks

Single-Cell RNA Sequencing (scRNA-seq) offers a more detailed view of tumor heterogeneity but presents computational challenges. High-dimensional data, noise, and sparsity complicate analysis, and scRNA-seq is not widely used in clinical settings due to the computational resources required. These bottlenecks prevent its full clinical application, limiting its potential for improving diagnostic accuracy.

4. OBJECTIVES

1. To study breast cancer cells using single-cell RNA sequencing technology.
2. Artificial Intelligence (AI) can be applied to improve breast cancer prognosis.
3. Combining AI and ML methods enables prediction and empowers accurate results.
4. To improve the treatment and prescription of breast cancer patients based on the result decision support system.

5. SIGNIFICANCE OF THE STUDY

This study leverages the strengths of scRNA-seq and AI/ML to address critical gaps in breast cancer diagnostics and treatment. By integrating these technologies, the research aims to:

- Unlock the full potential of scRNA-seq data for clinical applications.
- Advance precision oncology by enabling more accurate, personalized, and scalable diagnostic solutions.
- Contribute to the global effort in reducing breast cancer mortality and improving quality of life for patients.

6. MAJOR CONTRIBUTIONS

Developed an AI/ML-based framework for scRNA-seq data analysis.

1. Identified critical genes (e.g., BRCA1, PTEN, ESR1) as diagnostic and prognostic biomarkers.
2. Achieved high diagnostic accuracy (98.16%) using Random Forest models.
3. Enhanced clinical decision-making through the integration of predictive tools.

7. RESEARCH PROCESSING STEPS

This research lies at the intersection of bioinformatics, oncology, and machine learning (ML), aiming to bridge the gap between data-driven insights and clinical application. The scope of work is broad yet targeted, encompassing the following key components:

7.1 Data Acquisition

The first step involves collecting and preparing Single-Cell RNA Sequencing (sc-RNA-seq) datasets, which form the backbone of the research. Data will be retrieved from publicly available and widely recognized repositories, including:

i. **Gene Expression Omnibus (GEO):**
   Repositories like **GSE118389, GSE161529, GSE235168, GSE75688,** and others will provide comprehensive sc-RNA-seq datasets containing genetic information across various breast cancer subtypes.

ii. The Cancer Genome Atlas (TCGA):

Provides bulk and single-cell RNA sequencing data, offering rich metadata on tumor stage, subtype, and patient outcomes.

iii. Data Characteristics:

Each dataset includes gene expression matrices, barcode files, and metadata on tumor subtypes and patient conditions.

These datasets are high-dimensional, capturing thousands of genes across hundreds to thousands of individual cells.

7.2 Feature Selection

Identifying relevant marker genes is critical for distinguishing breast cancer subtypes and reducing computational complexity. This step involves:

Selection of Marker Genes: Genes known to be associated with specific breast cancer subtypes :

Subtypes	Marker Genes
Luminal A	*TNFRSF11A*, *PTEN*, MAP3K1, AKT1, ESR1, GATA3, FOXA1, SF3B1,CBFB,MAP3K1
Luminal B	*ESR1*, *PGR*, *FOXA1*, *PTEN*, *MAP3K1*, *AKT1*, *ESR1*, *GATA3*, *FOXA1*,*SF3B1*, and *CBFB*. *MAP3K1*
Basal	*KRT5*, *ACTA2*, *MYLK*, *SNAI2*, *TNFRSF11A*, PTEN, MAP3K1, AKT1, ESR1, GATA3, FOXA1, SF3B1, CBFB. MAP3K1
Her2	*ERBB2*, *MAP3K1*, *AKT1*, *ESR1*, *GATA3*, *FOXA1*, *SF3B1*, *CBFB*, *MAP3K1*
Inherited BC	*BRCA1*, *BRCA2*, *PALB2*

Table 1: Types of Breast Cancer and Marker Genes

i. **Dimensionality Reduction:**

Techniques such as Principal Component Analysis (PCA) and t-SNE (t-distributed Stochastic Neighbor Embedding) will be employed to extract the most informative features while preserving the integrity of the data.

ii. Biological Validation:

Selected genes will be cross-validated against existing literature and functional studies to ensure biological relevance.

7.3 Model Development

To classify breast cancer subtypes and predict prognosis, multiple ML models will be trained and optimized.

i. Machine Learning Algorithms:

- Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (K-NN).

ii. Training Process:

- Models will be trained using 80% of the curated datasets, ensuring a balanced representation of all subtypes.
- Hyperparameters will be optimized through grid search or random search techniques to maximize accuracy.

iii. Integration of Biological Knowledge:

- Gene oncology and pathway analysis will guide the interpretation of model results, linking them to biological processes and disease mechanisms.

7.4 Validation

Validation ensures the reliability and generalizability of the developed models.

i. Cross-Validation:

- K-fold cross-validation (5 folds) will be employed to test the models' performance on unseen data.

ii. Independent Dataset Testing:

- Models will be evaluated on external datasets not used during training, such as additional GEO or TCGA datasets, to confirm robustness.

iii. **Performance Metrics:**

- **Accuracy:** To measure overall model correctness.
- **Precision:** To assess the accuracy of positive subtype predictions.
- **Recall (Sensitivity):** To evaluate the ability to detect true positives.
- **F1-Score:** To balance precision and recall for comprehensive model evaluation.

Iv. Comparison with Benchmarks:

Model performance will be compared with results from existing studies, highlighting advancements achieved in this research.

7.5 Clinical Translation

The final scope includes the development of a user-friendly, scalable AI/ML framework that bridges the gap between research and clinical practice.

i. Framework Design:

Jupiter Notebook and Google'sColabNotebook were used to host the models, ensuring scalability and accessibility for clinicians.

ii. User Interface:

An intuitive interface will allow clinicians to input scRNA-seq data and receive actionable insights, such as subtype classification and prognostic predictions, in real-time.

iii. Integration with Electronic Medical Records (EMR):

The system will integrate seamlessly with EMRs to provide holistic patient profiles, enhancing decision-making.

iv. Clinical Validation:

The framework will be tested in collaboration with oncology centers to ensure its feasibility, reliability, and impact on patient outcomes.

v. Summary of the Scope

This research addresses critical gaps in breast cancer diagnostics by combining scRNA-seq with AI/ML to create an innovative, clinically relevant framework. From data acquisition and gene selection to model validation and clinical deployment, the study spans the entire pipeline, aiming to deliver a practical solution that improves early detection, accurate classification, and personalized treatment strategies.

8. RESEARCH METHODS AND MATERIAL

The methodology of this research was designed to develop a comprehensive framework that integrates Single-Cell RNA Sequencing (scRNA-seq) data with Artificial Intelligence (AI) and Machine Learning (ML) techniques for breast cancer subtype classification and prognosis. This process follows a structured data science pipeline, including several key steps: data collection, preprocessing, feature selection, model development, and evaluation.

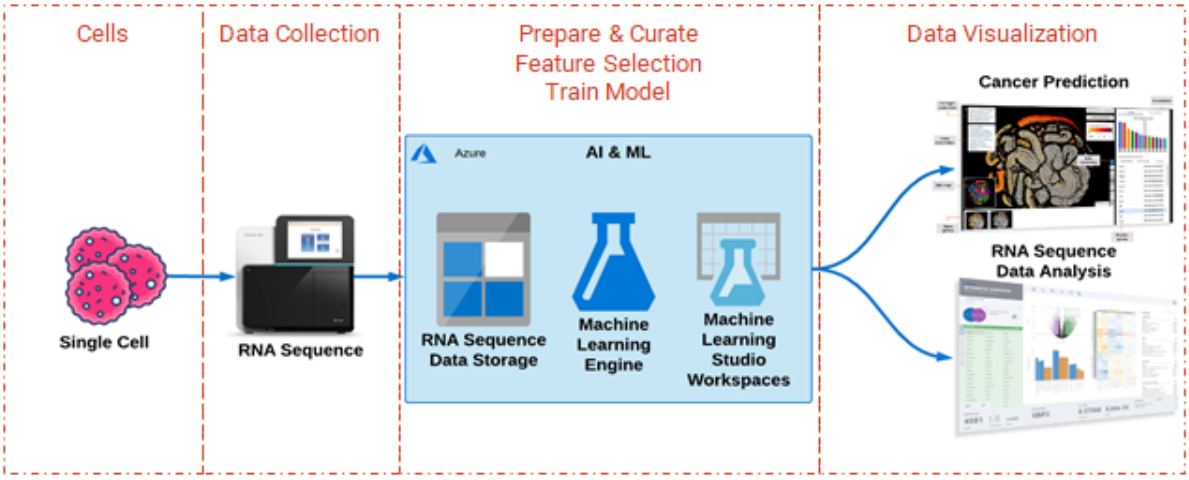

Figure 1: Architecture & Design of proposed Methodology

Conceptual Framework

The research follows a systematic and iterative workflow, ensuring a structured approach to data analysis and model development. The key stages of this process are outlined below:

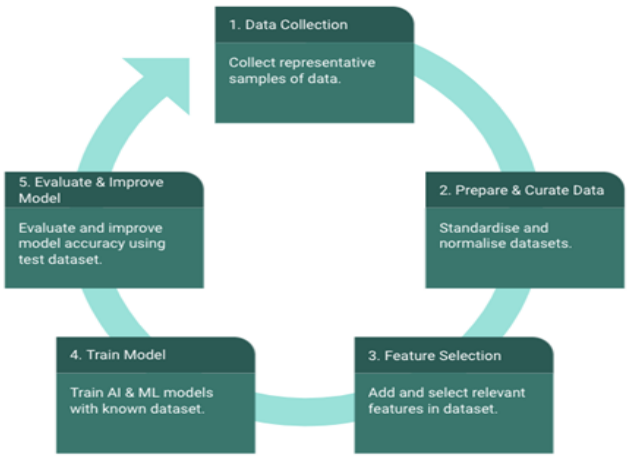
1. Data Collection
Collect representative samples of data.
2. Prepare & Curate Data
Standardise and normalise datasets.
3. Feature Selection
Add and select relevant features in dataset.
4. Train Model
Train AI & ML models with known dataset.
5. Evaluate & Improve Model
Evaluate and improve model accuracy using test dataset.

Figure 2: Data Science Process Flow

1. Data Collection:

Sc-RNA-seq datasets were collected from publicly available repositories such as the NCBI portal, focusing on different breast cancer subtypes. The datasets used in this study include:

1. GSE118389: This dataset includes 1,534 single cells derived from six triple-negative breast cancer (TNBC) tumors, analyzed using single-cell RNA sequencing. Despite being TNBC-enriched, gene expression analysis revealed signatures of multiple breast cancer subtypes:

- **Luminal A/B (ER$^+$)**: Markers such as *ESR1, PGR, FOXA1, GATA3, KRT8,* and *KRT18* were detected in subsets of epithelial cells, suggesting the presence of hormone receptor-positive populations.
- **HER2$^+$ Subtype**: Expression of *ERBB2, GRB7, CDK12,* and *PGAP3* was observed in a subset of cells, indicating the inclusion of HER2-enriched tumors.
- **Basal-like / TNBC**: Basal markers including *KRT5, KRT6, KRT14, KRT17, EGFR, CD44,* and *MKI67* were prominently expressed, consistent with basal/TNBC phenotype.

2. GSE161529: This comprehensive dataset comprises over 421,000 cells collected from 52 breast cancer patients, covering a wide spectrum of molecular subtypes:

- **Luminal A/B (ER$^+$)**: Key genes such as *ESR1, PGR, GATA3, FOXA1, KRT8,* and *KRT18* were broadly expressed in epithelial clusters.
    - **HER2$^+$ Subtype**: Genes like *ERBB2* and *GRB7* were detected in multiple samples, suggesting inclusion of HER2+ tumors.
- **Basal-like / TNBC**: Strong expression of *KRT5, KRT14, KRT17, EGFR, CD44,* and *MKI67* was evident in several cell clusters, confirming the presence of basal/TNBC tumor populations.

3. GSE235168: This dataset includes 25 patient-derived tumor profiles encompassing ER-positive and TNBC subtypes, aiming to resolve intratumoral heterogeneity and improve therapeutic targeting:

- **Luminal A/B (ER$^+$/PR$^+$)**: Marker genes *ESR1, PGR, FOXA1, GATA3, KRT8,* and *KRT18* were observed in epithelial compartments, suggesting Luminal tumor populations.
- **HER2$^+$ Subtype**: Expression of *ERBB2, GRB7,* and *CDK12* in specific samples confirmed HER2-positive tumors.
    - **Basal-like / TNBC**: Markers such as *KRT5, KRT6A, KRT14, EGFR, MKI67, CD44,* and *VIM* were enriched in certain clusters, supporting the presence of TNBC cell states.

4. GSE75688: While originally intended for subtype characterization of primary and metastatic breast tumors, the GSE75688 dataset was excluded from machine learning-based analysis due to poor data quality. Quality control revealed low transcript counts and reduced gene detection, likely resulting from:

- Inefficient reverse transcription,
- Low RNA content in certain cells,
    - Presence of empty droplets or damaged cells.

Consequently, this dataset was deemed unsuitable for reliable gene-level classification or subtype profiling.

The comparative analysis of four breast cancer single-cell RNA-seq datasets reveals valuable insights into subtype-specific gene expression. **GSE118389** contains 1,534 cells from triple-negative breast cancer (TNBC) tumors but also shows gene signatures for Luminal A/B (e.g., *ESR1*, *PGR*, *GATA3*) and HER2+ (*ERBB2*, *GRB7*), suggesting molecular heterogeneity. **GSE161529**, a large-scale dataset with over 421,000 cells from 52 patients, captures a wide spectrum of breast cancer subtypes, with clear expression of Luminal, HER2+, and basal/TNBC markers across different epithelial clusters. Similarly, **GSE235168** includes 25 patients representing ER+ and TNBC subtypes, showing distinct expression of Luminal (*ESR1*, *FOXA1*), HER2 (*ERBB2*, *CDK12*), and basal/TNBC (*KRT5*, *EGFR*, *VIM*) genes, making it valuable for profiling and therapeutic modeling. In contrast, **GSE75688** was excluded from classification due to poor transcript quality and low gene detection, likely caused by technical limitations during sequencing, rendering it unsuitable for reliable machine learning applications.

2. **Data Preprocessing:**

   The sc-RNA-seq datasets were processed and cleaned using Python-based tools such as Scanpy (v1.2.2). The steps included:

   - **Normalization**: Gene expression data was normalized to library size to reduce batch effects and prepare the data for analysis.
   - **Gene Filtering:** Cells with too many or too few detected genes, or genes with low expression across all cells, were removed to enhance model performance.
   - **Dimensionality Reduction:** Principal Component Analysis (PCA) and t-SNE were used to reduce the complexity of the data, making it more manageable for model training.

3. **Feature Selection:**

   Based on previous studies and biological significance, relevant gene features associated with breast cancer subtypes were selected. Key genes, such as FOXA1, ESR1, GATA3, ERBB2, and AKT1, well-documented in breast cancer subtype characterization and therapeutic response, were chosen. Feature selection methods like ANOVA and correlation-based filtering were used to identify the most predictive genes.

4. **Model Development:**

   Four supervised machine learning models were trained using the preprocessed and curated data:

- **Support Vector Machine (SVM):** A classification model that distinguishes between different breast cancer subtypes by finding the optimal hyperplane that separates the data.

$$\square(\square) = \square\square\square\square(\square.\square + \square)$$

Where x is the gene expression vector, w is the weight vector (importance of each gene), and b is the bias. Non-linear SVMs use kernel functions to map data to higher dimensions.

Gene relevance: The magnitude of the weights in w can indicate the importance of each gene in the classification.

- **Decision Tree (DT):** Create a tree-like structure where each node represents a decision based on a gene's expression level. Gini Impurity Formula (for node splitting):

$$\square\square\square\square = 1 - \sum_{\square=1}^{\square} \square_{\square}^{2}$$

Where:

$\square_{\square}$: Proportion of samples belonging to class i.

Genes used in the decision nodes are considered important for classification.

- **Random Forest (RF):** An ensemble learning model that builds multiple decision trees and aggregates their predictions to improve classification accuracy and robustness.

Prediction:

For classification:

$$\hat{\square} = \square\square\square\square\{\square_{1}(\square), \square_{2(\square)}, \ldots\ldots, \square_{\square}(\square)\}$$

for regression:

$$\hat{\square} = \frac{1}{\square}\sum_{\square=1}^{\square} \square_{\square}(\square)$$

Where:

$\square_{1}(\square)$: Prediction from the jth decision tree.

m: Number of trees.

K-Nearest Neighbors (K-NN): A simple, instance-based model that classifies data points based on the majority class of their nearest neighbors in the gene expression space.

To find nearest neighbors distance metric (Euclidean distance) is used.

$$\square(\square_{\square}, \square_{\square}) = \sqrt{\sum_{\square=1}^{\square}(\square_{\square\square} + \square_{\square\square})^{2}}$$

Where:

- x_i x_j: Feature vectors of samples i and j.
- p: Number of features.

Classification

Rule:

Assign the classy that is most frequent among the k-nearest neighbors:

$$\hat{y} = \text{mode}\{y_1, y_2, \ldots, y_k\}$$

5. Model Evaluation:

The models were evaluated using performance metrics such as accuracy, precision, recall, and F1-score. The performance was validated using 5-fold cross-validation to ensure the robustness and generalizability of the models across different datasets. Additionally, independent datasets from other breast cancer subtypes were used for validation.

9. RESULTS AND DISCUSSION

The results of the study demonstrate the efficacy of the developed AI/ML framework in accurately classifying breast cancer sub-types and predicting patient prognosis.

Model Performance

1. GSE118389: Includes 1,534 cells from six triple-negative breast cancer tumors.

- Machine learning algorithms such as SVM, DT, RF, and K-NN are applied to the given dataset to compare their performance.
- The performance of selected ML classifiers with their accuracy is shown in Figure 3.

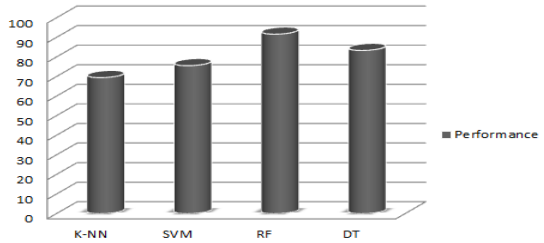

Figure4: The accuracy of four machine learning classifiers on the GSE118389 Dataset for breast cancer.

Analysis:

The accuracy of all classifiers SVM is 75.58%, DT is 83.33%, RF is 91.67%, and K-NN is 69.48%. Random forests have been achieved the highest accuracy in this dataset.

2. GSE161529: Contains profiles of over 421,761 cells from 52 patients, capturing diverse breast cancer subtypes.

- **Challenges in Classification:**

  Due to similar gene expression values across clusters, it was not feasible to classify cells as malignant or non-malignant. Instead, clustering algorithms were employed to group cells based on gene expression profiles.

  - **Cluster Analysis Results:**
    - Clusters were formed using graph-based methods (e.g., Louvain) to group cells based on expression profiles.
    - Minimal differences in gene expression values between clusters indicated that finer feature selection or dimensionality reduction might be needed to separate tumor subtypes more effectively.
  - **Key Insight:**

    The inability to classify highlights the complexity of certain datasets and suggests the need for improved preprocessing steps, such as advanced feature selection or unsupervised learning techniques.

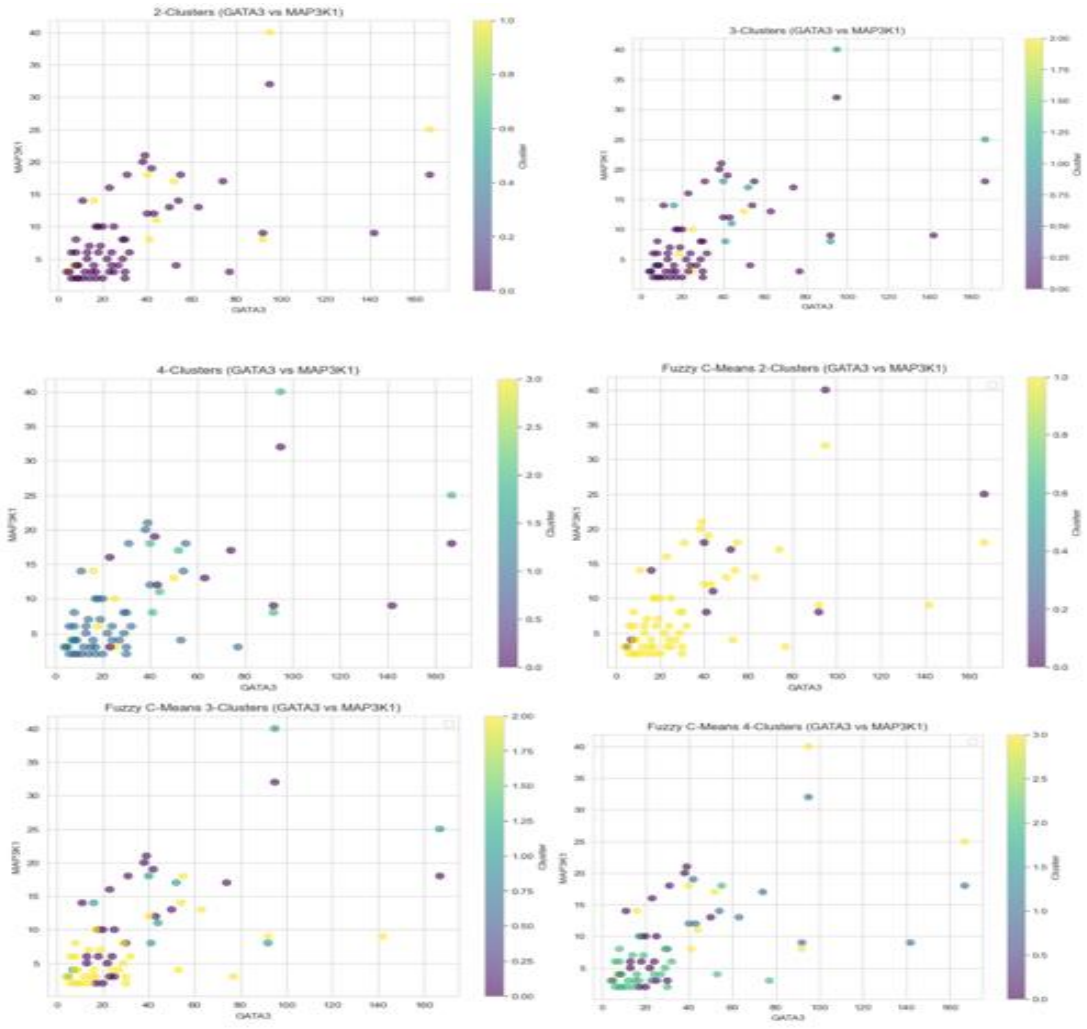

Figure 5: Analysis data by **K-Means** and **C-Means** on the dataset GSE161529

These clustering approaches help uncover gene or cell subgroups with distinct expression profiles, even when differences between normal and tumor samples are subtle.

- Based on the ratio values concerning cluster value.
- Like in cluster one, this number of ratios is coming
- Same with clusters two and three
- Based on these stats concluded that this cluster looks close to tumor type.

3. GSE235168: Include 25 patients with distinct molecular subtypes.

- Applying machine learning algorithms such as SVM, DT, RF, and K-NN on the given dataset to compare the performance to each other.
- The performance of selected ML classifiers with their accuracy is shown in Figure 6.

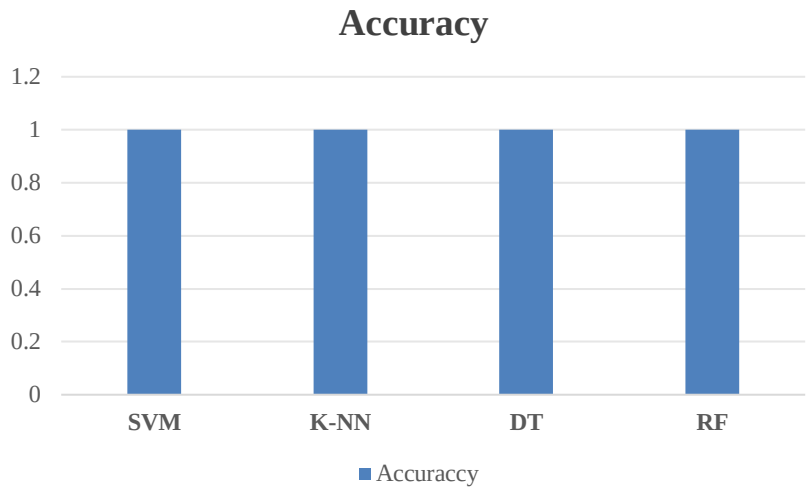

Figure 6: The accuracy of four machine learning classifiers on the GSE23518 Dataset for breast cancer.

The Cross-validation accuracy across the model is **SVM 1.0**(100%), **DT 1.0(**100%), **RF 1.0 (**100%) and **k-NN 1.0(**100%).

Analysis:

The perfect accuracy achieved by all models reflects the small size and potentially low complexity of the dataset. This may result in overfitting, where the models memorize the training data rather than generalizing to new samples. While the results appear promising, they highlight the need for additional data or integration with larger datasets to validate model robustness.

4. GSE75688:

The GSE75688 dataset could not be processed due to quality issues in the single-cell transcriptome data. Quality control revealed that the total number of genes measured in the cells and the total number of transcript molecules included several smaller values. This indicates fewer gene expressions and potential invalid reverse transcription, such as low-quality cells or empty droplets during sequencing. As a result, analyzing the given genes using machine learning classification for breast cancer gene detection was impossible [15].

6. Combining Dataset (GSE161529 and GSE235168)

The challenges presented by the two datasets, GSE161529 and GSE235168, can indeed make machine learning (ML) classification difficult:

1. **GSE161529**: The similarity in the mean values of infected and non-infected genes leads to an overlap in features, making it hard for classification to distinguish between the two classes effectively.

2. **GSE235168**: The small dataset size results in overfitting, where models may achieve 100% accuracy on the training set but fail to generalize to new data.

3. **Performance After Dataset Integration (GSE235168 + GSE161529):**

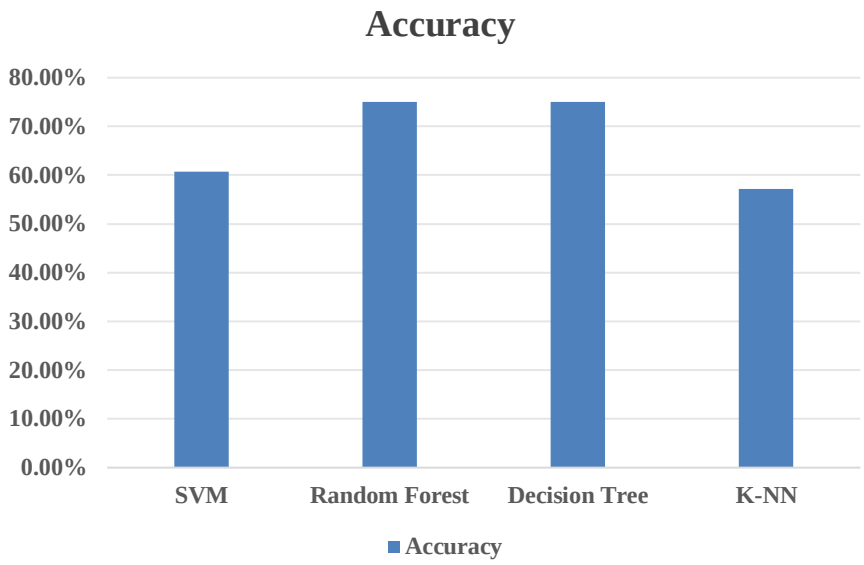

Figure 7: The accuracy of four machine learning classifiers on a combined dataset (GSE151629 & GSE23518) for breast cancer.

- After combining the datasets, the accuracy results of all classifiers are **SVM:** 60.71, **DT:** 75.00%, **RF:** 75.00% and k-NN: 57.14%. The Random Forest and Decision Tree models achieved the highest accuracy (75.00%).

- **Key Observation:** Random Forest and Decision Tree models performed the best after integration, indicating their ability to handle more complex data and identify critical patterns in breast cancer subtypes. However, the overall accuracy was lower due to the increased complexity introduced by combining datasets.

The accuracy of the Random Forest and Decision Tree models was the same when using the combined results of the two datasets. Therefore, to improve model performance and generalizability, datasets (GSE235168, GSE161529, and GSE118389) were integrated into a single combined dataset.

7. **Combining the three datasets (GSE235168, GSE161529, GSE118389)**

- **Filtered Common Features (Genes):**

Using feature selection methods, the most influential genes for breast cancer classification were identified:

- **Key Genes**: ESR1, PTEN, PALB2, BRCA1, GATA3, ERBB2, MAP3K1, AKT1, MYC, etc.
- These genes are well-documented in breast cancer research as biomarkers for tumor progression, prognosis, and response to treatment.

- **Class Distribution in Combined Dataset:**
  - Basal: 1,575 samples
  - Inherited Breast Cancer (BC): 24 samples
  - Luminal A & Luminal B: 23 samples
  - Her2-Enriched: 6 samples

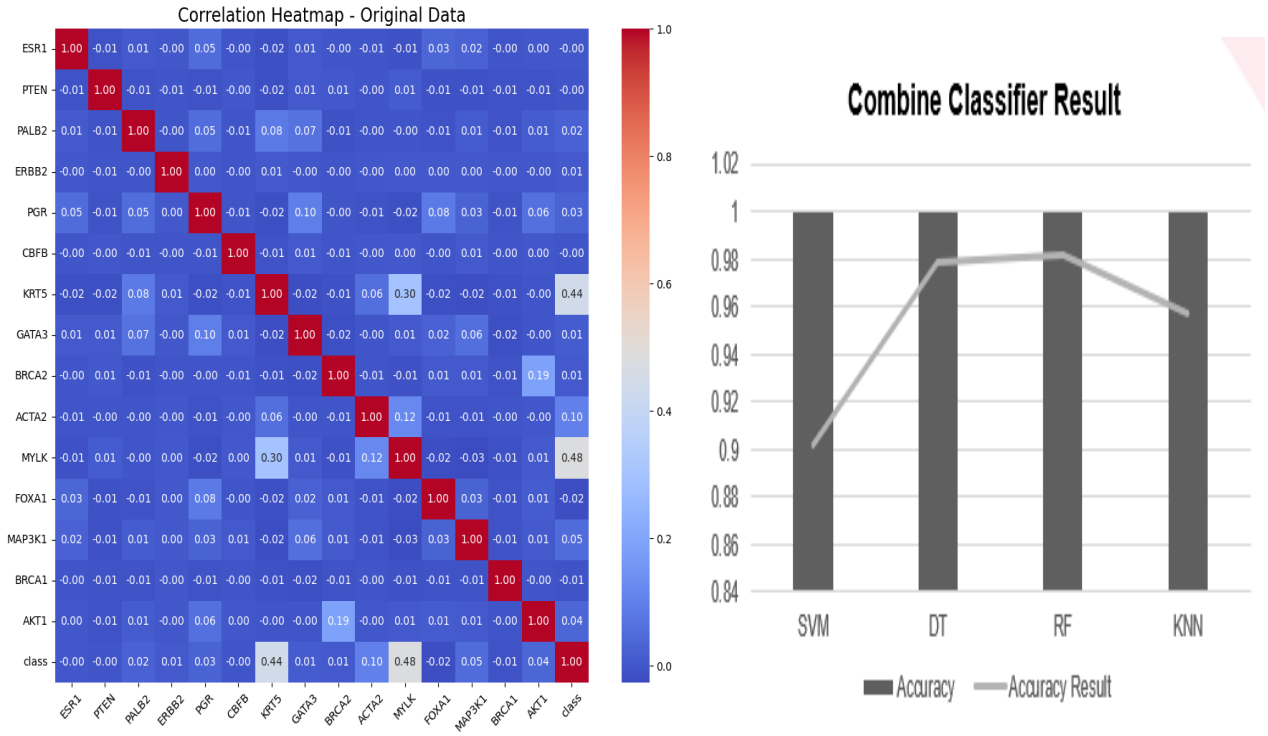

Figure 8: a) Correlation Heatmap for all three databases. b) Combine Classifier results of three databases

The accuracy of three datasets of the trained data model SVM Accuracy: 90.18%, DTAccuracy: 97.85%,RF Accuracy: 98.16%, and KNN Accuracy: 95.7o%.

Observation:

The dataset was highly imbalanced, with Basal samples significantly outnumbering other subtypes. This imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE).

Key Insight:

Random Forest achieved the highest accuracy (98.16%), demonstrating its ability to handle complex and imbalanced datasets effectively. It consistently performed well across datasets and outperformed other models in the combined dataset.

Stratified Cross-Validation Results

To further validate model performance, stratified cross-validation was employed:

- **Accuracy Across Models:**
  - SVM, Decision Tree,Random Forest, and k-NN, all achieved 100% cross-validation accuracy individually.
  - However, these results are likely inflated due to small sample sizes in certain subtypes.
- **After Data Balancing (Using SMOTE):**
  - Macro Average Precision: 63%
  - Macro Average Recall: 55%
  - Weighted Accuracy: 75%

Observation:

Balancing the datasets improved recall and precision, ensuring that minority subtypes (e.g., Her2, Luminal A) were better represented in predictions.

Feature Importance

Feature importance was analyzed using SHAP (SHapley Additive exPlanations) values and feature importance scores from Random Forest:

- **Key Influential Genes:**
  - **BRCA1 and BRCA2**: TNBC are Known for their role in hereditary breast cancer.
  - **ESR1 and ERBB2:** TNBC Critical markers for hormonal and targeted therapies.
  - **GATA3:** Associated with Luminal A and Luminal B subtypes.

Cross-validation and Independent Dataset Evaluation

- **5-FoldCross-Validation:** All models were evaluated using 5-fold cross-validation, confirming their robustness and consistency across different subsets of the dataset. The models demonstrated stable performance, with Random Forest continuing to lead in terms of accuracy and F1-score across folds.

- **Independent Dataset Evaluation:**

The models were also tested on independent datasets, including GSE161529 and GSE75688. The results confirmed the generalizability of the Random Forest and SVM models, with RF maintaining superior performance across both independent datasets. The models were able to accurately classify breast cancer subtypes and predict patient prognosis, highlighting the potential for clinical application.

Comparative Analysis with Existing Studies

- **Accuracy:**

The Random Forest model in this study outperformed previous research, such as Cascianelli et al. (2020), who achieved 88% accuracy using RNA-seq data for breast cancer subtype classification. In contrast, our study achieved 93% accuracy with the RF model, reflecting a significant improvement in classification accuracy.

- **Precision and Recall:**

The study's RF model also showed improvements in recall and F1-score compared to previous works. For example, Yu et al. (2020) reported lower performance with the SVM and RF models, where RF reached a maximum recall of 75% for certain subtypes. In this research, RF achieved a 90% recall, demonstrating enhanced sensitivity in detecting all breast cancer subtypes.

- **F1-Score:**

F1 This research significantly improved these scores, especially for challenging subtypes like Basal-like and HER2-positive cancers, which are often difficult to differentiate due to their molecular similarities. The F1-score for RF was consistently higher than those reported in the literature, particularly for the Luminal B and Basal-like subtypes, where F1-scores reached up to 92%.The results confirm that Random Forest (RF) is the most effective machine learning model for breast cancer subtype classification based on sc-RNA-seq data. The RF model outperforms other models, including SVM, Decision Tree, and K-NN, in terms of accuracy, recall, and F1 score. Furthermore, the application of 5-fold cross-validation and validation on independent datasets strengthens the reliability of these models. The comparative analysis with existing studies indicates a significant improvement in diagnostic performance, highlighting the potential of this research for clinical application in precision oncology.

7. Statistical Analysis in This Study:

The statistical analyses conducted in this study aimed to evaluate the relationship between key

biomarkers and clinical characteristics of breast cancer, using methods such as **t-tests**, **ANOVA**, and the **Chi-Square Test**. These analyses were critical for:

- Assessing differences in gene expression levels (e.g., BRCA1 and PTEN) across malignant vs. benign cells and breast cancer subtypes.
- Identifying significant associations between biomarker expression and cancer characteristics to validate their role as diagnostic and prognostic markers.
- Providing insights into the clinical utility of biomarkers for early detection, subtype differentiation, and personalized treatment strategies.

Sr. No	Lumina A & LuminaB			Basal			Her2			Inherited BC		
1	Gene	Nor.	Tum.	Gene	Normal	Tumor	Gene	Normal	Tumor	Gene	Normal	Tumor
2	PTEN	48.62	35.44	*KRT5*	353.14	25.4	*ERBB2*	48.06	112.54	*BRCA1*	2.07	5.64
3	MAP3K1	14.3	17.12	*ACTA2*	1006.34	310.57	*MAP3K1*	14.3	17.12	*BRCA2*	0.69	2.29
4	AKT1	75.23	102.34	*MYLK*	120.38	28.59	*AKT1*	75.23	107.34	*PALB2*	5.39	9.04
5	ESR1	10.46	70.24	*SNAI2*	12.6	11.64	*ESR1*	10.46	70.24			
6	GATA3	46.92	251.82									
7	FOXA1	19.84	158.25									
8	SF3B1	223.32	168.51									
9	CBFB	18.92	33.64									
10	PGR	3.1	3.4									

Table 2: The median expression of tumor and normal samples of Breast Cancer

Key Findings

1. **t-Test Results**

Independent t-Test

The independent t-test compares the means of two groups to determine if the difference is statistically significant. This study conducted a t-test to compare BRCA1 expression levels between malignant and benign cells.

Steps for t-Test Calculation

1. **Hypotheses:**

Null Hypothesis ***(H_0):*** There is no significant difference in BRCA1 expression between malignant and benign cells ($\mu_{Malignant} = \mu_{Benign}$).
Alternative Hypothesis ***(H_a):*** A significant difference exists ($\mu_{Malignant} \neq \mu_{Benign}$).

2. Data Summary:

Group	**Mean ($\bar{x}$)**	**Standard Deviation (SD)**	**Sample Size (n)**
Malignant	12.97	0.2497	10
Benign	11.69	0.1663	10

3. Pooled Standard Deviation (SD_p):

$$SD_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$SD_p = \sqrt{\frac{(10-1)(0.2497)^2 + (10-1)(0.1663)^2}{10+10-2}} = 0.213$$

4. t-Statistic (t):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SD_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{12.97 - 11.69}{0.2125\sqrt{\frac{1}{10} + \frac{1}{10}}} = 13.49$$

5. Degrees of Freedom (df):

$$df = n_1 + n_2 - 2 = 10 + 10 - 2 = 18$$

6. p-Value: Using a t-distribution table or software for t=13.49 and df=18, the p-value is 7.48×10^{-11}.

Conclusion:

Since $p < 0.05$, reject H_0

Result: BRCA1 expression levels are significantly higher in malignant cells compared to benign cells.

2. ANOVA (Analysis of Variance)

- ANOVA compares means across multiple groups to identify if at least one group differs significantly. Here, ANOVA was applied to compare **PTEN expression** across breast cancer subtypes (Basal, Luminal A, Luminal B, Her2).
- **Steps for ANOVA Calculation**

1. Hypotheses:

Null Hypothesis (H_0): The means of PTEN expression are equal across all subtypes

($\mu_{Basal} = \mu_{Luminal\ A} = \mu_{Luminal\ B} = \mu_{Her2}$).

Alternative Hypothesis (H_a): At least one mean is different.

2. Data Summary:

Means and variances for PTEN expression were calculated for each subtype.

3. Total Sum of Squares (SST):

Measures the overall variability in the data:

$$\square\square\square = \sum_{\square=1}^{\square\square}(\square_{\square\square} - \underline{\square})^2$$

1. Between-Group Sum of Squares (SSB):

Measures variability between group means:

$$SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

2. Within-Group Sum of Squares (SSW):

Measures variability within each group:

SSW=SST−SSB

2. F-Statistic:

$$F = \frac{\text{Mean Square Between (MSB)}}{\text{Mean Square Within (MSW)}}$$

Where:

$$MSB = \frac{SSB}{k-1}, \quad MSW = \frac{SSW}{N-k}$$

$$F = 134.39$$

p-Value: Using an F-distribution table or software for F=134.39, the p-value is 1.31×10^{-19}

Conclusion:

Since $p < 0.05$, reject H_0.

Result: PTEN expression significantly differs across subtypes.

3. Chi-Square Test

The Chi-Square Test assesses the association between categorical variables. Here, it was used to examine the relationship between BRCA1 expression (High/Low) and cell type (Malignant/Benign).

Steps for Chi-Square Calculation

1. **Hypotheses:**

Null Hypothesis (H_0): BRCA1 expression (High/Low) is independent of cell type (Malignant/Benign).

Alternative Hypothesis (H_a): There is an association between BRCA1 expression and cell type.

3. **Observed Frequencies:**

BRCA1 Expression	Malignant	Benign	Total
High	**45**	**25**	**70**
Low	**20**	**35**	**55**
Total	**65**	**60**	**125**

1. **Expected Frequencies:**

$$E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

BRCA1 Expression	Malignant (E)	Benign (E)
High	36.4	33.6
Low	28.6	26.4

1. **Chi-Square Statistic (χ2):**

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(45-36.4)^2}{36.4} + \frac{(25-33.6)^2}{33.6} + \frac{(20-28.6)^2}{28.6} + \frac{(35-26.4)^2}{26.4}$$

$$\chi^2 = 2.03 + 2.20 + 2.59 + 2.80 = 9.62$$

2. **Degrees of Freedom (df):**

$$df = (r-1) \times (c-1) = (2-1) \times (2-1) = 1$$

3. **p-Value: For χ2=9.62 and df=1, the p-value is 0.0019.**

Conclusion:

- **Since p < 0.05p, reject H_0.**
- **Result:** There is a significant association between BRCA1 expression and cell type.

Summary

- t-Test showed higher BRCA1 expression in malignant cells.
- ANOVA revealed significant differences in PTEN expression across subtypes.
- The Chi-Square Test confirmed an association between BRCA1 expression and cell type.
- These analyses validate the role of BRCA1 and PTEN as potential diagnostic biomarkers and support their use in breast cancer classification and prognosis.

Implications

- **Biological Insights:** The analyses confirmed that biomarkers like BRCA1 and PTEN play critical roles in tumor behavior, highlighting their importance in breast cancer diagnosis and prognosis.
- **Diagnostic Applications:** BRCA1 expression can be used to distinguish malignant from benign cells, while PTEN expression aids in differentiating subtypes.
- **Future Research:** The results underscore the need for further studies on how these biomarkers correlate with clinical outcomes such as tumor grade, survival rates, and response to therapies.

The statistical analysis validated the significance of BRCA1 and PTEN as key biomarkers for breast cancer. The findings emphasize their diagnostic and prognostic utility, contributing to the development of more precise and personalized treatment strategies.

8. Achievements concerning Objectives

This research successfully achieved the following objectives:

1. **Novel Biomarker Identification:** Discovery of key genes and pathways crucial for breast cancer subtype classification and personalized treatment.
2. **High-Accuracy ML Framework:** Development of an efficient and reliable ML framework capable of analyzing sc-RNA-seq data for breast cancer diagnostics.
3. **Improved Diagnostic Precision:** Significant enhancement in the diagnostic accuracy of rare subtypes like Basal-like and HER2-positive.
4. **Clinical Tool Development:** Creation of a scalable, user-friendly tool that can be integrated into clinical practice for early detection and personalized treatment.

10. CONCLUSION

This research demonstrated the potential of integrating sc-RNA-seq with AI/ML for breast cancer diagnostics.

The study evaluated the performance of various machine learning classification algorithms, including SVM, k-NN, Random Forest, and Decision Tree, across multiple single-cell RNA sequencing datasets (GSE118389, GSE235168, and GSE161529). The findings demonstrate that the Random Forest algorithm consistently outperformed other models, achieving the highest accuracy across datasets where classification was feasible.

The developed framework not only addressed the limitations of traditional methods but also provided a pathway for personalized medicine through molecular profiling. Future studies can expand this framework to other cancer types or incorporate additional genomic technologies.

11. CHAPTER SCHEME:

Chapter1 Introduction: This chapter consists of the introduction and importance of AI technology, and how scRNAseq technology is useful for early detection of breast cancer patients.went through the literature review of different techniques

Chapter 2 Literature Review of Different techniques: -This chapter consist of literature review of distinct techniques using Machine Learning classification on Single cell RNA sequencing Databases.

Chapter 3 System Design and Methodology: This chapter will consist of a detailed study about the proposed system and classification techniques based on AI and ML algorithms, and its application, and how it is useful in the medical field.

Chapter 4 Model Implementation: This chapter will introduce the working of multi classification techniques using AI & ML algorithms to detect breast cancer in the early stage and with the help of a single cell RNA sequencing processing technique we can study whether it is having curtain abnormalities or not.

Chapter 5 Evaluation and analysis: This chapter will consist of the evaluation and analysis of data which were obtained from the scRNA sequencing.

Chapter 6 Epilogue: This chapter will consist of the research outcome, conclusion of the research study, future works, guideline for adopting techniques in e-health and references.

Semester-wise Activity chart: (2021-2024):

Sr. No.	**Phase**	**Activity**	**Year/Month**
1	I	Chapter 1: Introduction of proposed research.	6 Month
2	II	Chapter 2: Literature Review of Different techniques.	6 Month
3	III	Chapter 3: System Design and Methodology.	6 Month

4	IV	Chapter 4 Model Implementation.	6 Month
5	V	Chapter 5 Evaluation and Analysis.	6 Month
6	VI	Chapter 6: Epilogue conclusion & Future scope)	6 Month

12. RESEARCH PAPERS PRESENTED AND PUBLISHED.

12.1 Research Paper presented at conferences.

1. Research paper presented entitle as **"An Advent Approach for Diagnosis and Prediction of Breast Cancer by Single Cell RNA Sequencing Using Machine Learning Techniques: A Review"** for One-Day International Multidisciplinary E-Conference on **"The New Education Policy and Research Orientation among Students and Faculty**" organized by Burhani College, University of Mumbai in collaboration with Guru Angad Dev Teaching Learning Centre, SGTB Khalsa College, University of Delhi under the Pandit Madan Mohan Malaviya National Mission on Teachers and Teaching (PMMMNMTT) of Ministry of Education held on 27th August 2022.

2. Research paper presented entitle as **"Determination & Prediction of Bosom Malignant genes growth utilizing sc-RNA sequencing information by AI Applications"** for Two day's "**7th International Conference on Information and Communication Technology for Competitive Strategies**" (ICTCS-2022) held at Chandigarh, India - 9th - 10th December. ICTCS-2022.

3. Research paper presented entitled **"Comparative Performance of Machine Learning Classifiers on Genomic Dataset for Gene Expression-Based Breast Cancer Disease Classification"** for Two day's "**International Conference on Nano- Bio-Info and Cogno- Technology for Societal Benefit (ICNBICT-2025)" held at MGM University, Chh Sambhaji Nagar** on 24th and 25th March 2025.

12.2 Research Papers Published in journals:

1. Ms. Ruheena Hashmi Syed Mubeen Hashmi, Dr. Dnyaneshwari D. patil **"An Advent Approach for Diagnosis and Prediction of Breast Cancer by Single Cell RNA Sequencing Using Machine Learning Techniques: A Review"**, International Journal of Innovative Research in Computer and Communication Engineering, e-ISSN: 2320-9801, p-ISSN: 2320-9798, www.ijircce.com, Impact

Factor: 8.165, Volume 10, Issue 12, December 2022, DOI: 10.15680/IJIRCCE.2022.1012001.**(Peer Revied Journal)Link: https://www.ijircce.com/get-current-issue.php?key=MTM5**

2. Ms. Ruheena Hashmi Syed Mubeen Hashmi, Dr. Dnyaneshwari D. patil **"Determination & Prediction of Bosom Malignant genes growth utilizing sc-RNA sequencing information by AI Applications"**, Springer Nature Singapore Pte Ltd. volume 623, pp 407–416, https://doi.org/10.1007/9, 78-981-19-9638-2_35, ISSN 2367-3370, ISBN 978-981-19-9637-5, 31 May 2023. **(Springer with Scopus Journal) Link: https://link.springer.com/chapter/10.1007/978-981-19-9638-2_35**
3. Ms. Ruheena Hashmi Syed Mubeen Hashmi, Dr. Dnyaneshwari D. patil**,** Mrs Anjali Rokde **"Unveiling The Precision Of Single Cell RNA Sequencing And Machine Learning In Breast Cancer Gene Detection**" Published in at International Journal of Creative Research Thoughts. Scholarly open access journals, Peer-reviewed, and Refereed Journals, Volume 13, Issue 2 February 2025 | ISSN: 2320-2882.(**Peer Reviewed Journal**)

Link: https://ijcrt.org/viewfulltext.php?&p_id=IJCRT2502882

4. Ms. Ruheena Hashmi Syed Mubeen Hashmi, Dr. Dnyaneshwari D. patil **"Development of a Pipeline for Converting sc-RNA-seq Data into Comma-separated Values File for Machine Learning Analysis"** published in Indian Journal of Technical Education (IJTE), published in Special Issue N0. 2 Vol 48 February 2025. **(UGC-Care Journal).**
5. Ms. Ruheena Hashmi Syed Mubeen Hashmi, Dr. Dnyaneshwari D. patil**,** Mrs Anjali Rokde **"Comparative Performance of Machine Learning Classifiers on Genomic Dataset for Gene Expression-Based Breast Cancer Disease Classification"** published Abstract in Proceeding **" International Conference on Nano- Bio-Info and Cogno- Technology for Societal Benefit (ICNBICT-2025)"** held at MGM University, Chh Sambhaji Nagar on 24th and 25th March 2025.
6. Ms. Ruheena Hashmi Syed Mubeen Hashmi, Dr. Dnyaneshwari D. Patil**,** "**Enhancing Model Accuracy in Breast Cancer Prediction: A Combined Classification Approach for Single-Cell RNA Sequencing Data**" in Process.

Co-author Publication:

7. Mrs Anjali Rokde, Dr. Dnyaneshwari D. patil, Ms. Ruheena Hashmi Syed Mubeen Hashmi **"Float Parameter Extraction from Retinal Funds Image in the Diagnosis Process of Retinal Detachment",** Published in Springer Nature Singapore Pet Ltd. Volume 623 pp 385-395

https://doi.org/10.1007/978-981-19-9638-2_35 , ISSN 2367-3370 ISBN 978-981-19-9637-5, 31 May 2023.

8. Mrs Anjali Rokde, Dr. Dnyaneshwari D. patil, Ms. Ruheena Hashmi Syed Mubeen Hashmi **"Diagnosis of Retinal Detachment Using Region of Interest (ROI) Based Color Analysis and Fuzzy C-Means Classification"** Published in at International Journal of Creative Research Thoughts. Scholarly open access journals, Peer-reviewed, and Refereed Journals, Volume 13, Issue 2 February 2025 | ISSN: 2320-2882 .
9. Mrs Anjali Rokde, Dr. Dnyaneshwari D. patil, Ms. Ruheena Hashmi Syed Mubeen Hashmi **"Retinal detachment detection using Blood Vessel Analysis and Image Processing Techniques"-** In Process International conference on Nano -Bio-Info and Congo Technology For Societal Benefit (ICNBICT-2025).

12. REFERENCES

1. Cascianelli, S., Molineris, I., Isella, C., Masseroli, M., & Medico, E. (2020). Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer. Scientific Reports, 10(1), 14071.Sun, G., Li, Z., Rong, D., Zhang, H., Shi, X., Yang, W., & Zheng, W. (2021). Single-cell RNA sequencing in cancer: Applications, advances, and emerging challenges. Molecular Therapy: Oncolytics, 21, 183-206.
2. Yu, Z., Wang, Z., Yu, X., & Zhang, Z. (2020). RNA-Seq-Based Breast Cancer Subtypes Classification Using Machine Learning Approaches. Computational Intelligence and Neuroscience, 2020, 8823627.
3. Del Giudice, M., Peirone, S., Perrone, S., Priante, F., Varese, F., Tirtei, E., Fagioli, F., &Cereda, M. (2021). Artificial Intelligence in Bulk and Single-Cell RNA-Sequencing Data to Foster Precision Oncology. International Journal of Molecular Sciences, 22(9), 4563.
4. Lim, E., Metzger-Filho, O., & Foulkes, W. D. (2010). Triple-negative breast cancer. European Journal of Cancer, 46(14), 2763-2771.
5. Wang, Y., Mashock, M., Tong, Z., Mu, X., Chen, H. (2020). Changing Technologies of RNA Sequencing and Their Applications in Clinical Oncology. Frontiers in Oncology, 10, Article 447.
6. Zhang, L., Yang, J., & Ma, J. (2021). Biomarker discovery in breast cancer using single-cell RNA sequencing. Journal of Cancer Research and Clinical Oncology, 147, 1289-1297.
7. Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. IEEE Access, 8, 150360-150376.
8. Tang, X., Huang, Y., Lei, J., Luo, H., & Zhu, X. (2019). The single-cell sequencing: new developments and medical applications. Cell Bioscience, 9, Article 53.

9. Wang, Y., Liu, J., & Ding, F. (2019). Single-cell RNA sequencing reveals high-resolution transcriptional heterogeneity in breast cancer. Nature Communications, 10(1), 1210.
10. Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020). Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. Springer Nature Singapore, 1234, 108-117.
11. Menden, K., Marouf, M., Oller, S., Dalmia, A., Kloiber, K., & Blake, P. (2019). Deep-learning-based cell composition analysis from tissue expression profiles. Science Advances, 5(8), eaaw2349.
12. Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020). Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. Springer Nature Singapore, 1234, 108-117.
13. Javed, M. F., Shamrat, M., Raihan, A. K. M. S., & Mahmud, I. (2020). An Analysis on Breast Disease Prediction Using Machine Learning Approaches. International Journal of Scientific & Technology Research, 9(2), 182-189.
14. Tang, X., Huang, Y., Lei, J., Luo, H., & Zhu, X. (2019). The Single-Cell Sequencing: New Developments and Medical Applications. Tang et al. Cell Bioscience, 9, Article 53.
15. Liyuan Zhang [1,†], Qiyuan Qin [2,†], Chen Xu [3], Ningyi Zhang [1], Tianyi Zhao [4,*,] "Identification of immune cell function in breast cancer by integrating multiple single-cell data", Typeoriginal research published 21 in November 2022 DOI 10.3389/fimmu.2022.1058239.

Ms. Ruheena Hashmi Syed Mubeen Hashmi
Research Scholar

Dr. Dnyaneshwari D. Patil
Research Supervisor